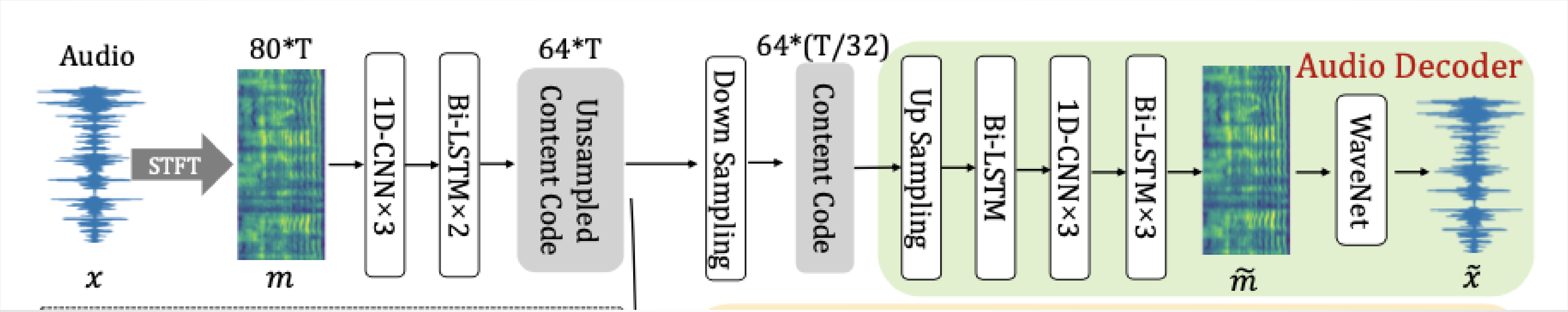
清華大學
Tsinghua University

Enhanced exemplar autoencoder with cycle consistency loss in any-to-one voice conversion

Weida Liang
2022.4.11

Contents

Exemplar Autoencoder



Encoder **Decoder** **Vocoder**

Kangle Deng, Aayush Bansal, Deva Ramanan, “UNSUPERVISED AUDIOVISUAL SYNTHESIS VIA EXEMPLAR AUTOENCODERS“ in ICLR 2021

Compressibility of Audio Speech

- Speech contains two types of information: $x = f(s, w)$
  - (i) content(large variance) (ii) style(little variance)

- Human Acoustics:
  - $Error\big(f(s_1, w_0), f(s_2, w_0)\big) \leq Error\big(f(s_1, w_0), f(s_2, w)\big), \forall w \in W$

- Autoencoder for Style Transfer:
  - $D\big(E(\hat{x})\big) \approx argMin_{t \in M} Error(t, \hat{x})= argMin_{t \in M} Error(t, f(s_1, w)) \approx f(s_2, w)$
    - M is the manifold spanning a particular style s_2.
  - Given sufficiently small bottlenecks, autoencoders can project out-of-sample points into the input subspace, so as to minimize the reconstruction error of the output.

Properties

- Pros
  - A simple autoencoder framework(CNN+BI-LSTM)
  - Data-efficient and few-shot
    - given a target speech with a particular style, learn an autoencoder specific to that target speech
- Cons
  - Bad performance on cross-gender task
    - the content from the bottleneck and the speaker style from the weights are not purely factorized.

Tsinghua University

Timeline

Date	Work
2021.7～2021.8	Finish baseline
2021.9～2021.10	Finish Cycle loss Model
2021.11	Design a project website Do GOP tests
2021.12	Finish a first draft of paper Add never-before-seen tests
2022.1	Wav2vec model configuration and training
2022.1～2022.2	Add CycleVAE comparison
2022.2～2022.3	Finish all experiments and write paper
2022.3～2022.4	Submit paper and add supplementation tests

Timeline

2021.7～2021.8.12 bi-weekly report， finish Exemplar Autoencoder baseline
2021.8.20～2021.9.13 two possible plans

Single Autoencoder
With multiple speakers

Multiple Autoencoders
With multiple speakers

Alternative solutions

To improve the information disentangled capacity of exemplar autoencoder, we design two alternative training methods.

One
a. Train the autoencoder with an arbitrarily large number of speakers.
In this stage, we assume that the variance of speech content is larger than speaker style,
so the bottleneck contains more content information.

b. Fix the encoder, and finetune the decoder using speech from a target speaker, and learn a specific exemplar autoencoder.
This stage is used to capture more speaker-specific information.

Two
a. Train N exemplar autoencoders with speech from N speakers.

b. Fix all the decoders of the N exemplar autoencoders, and then train one speaker-shared encoder.
By this way, we can squeeze the speaker-irrelevant content information into the encoder.

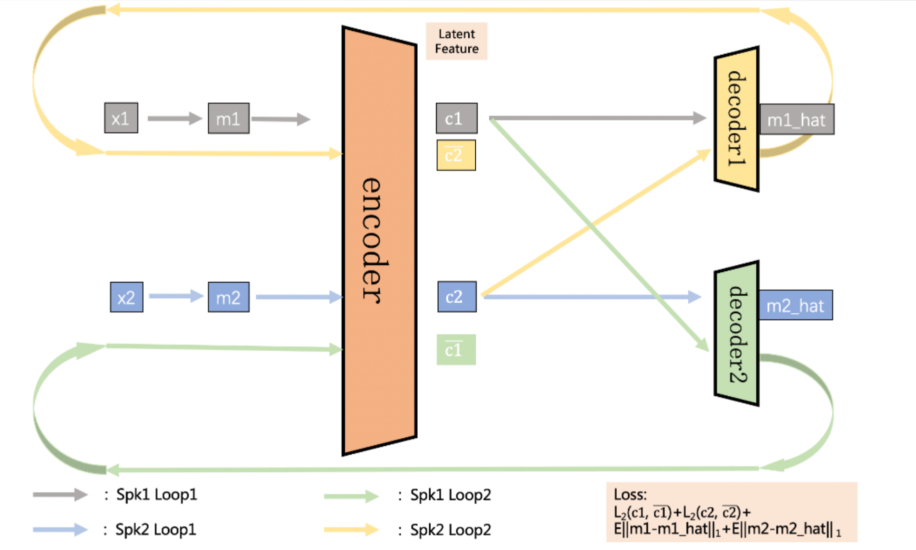
c. Fix the encoder, and train the decoder using speech from a target speaker, and learn a specific exemplar autoencoder.
This stage is used to capture speaker style.

The project website has been updated at http://166.111.134.19:7777/liangwd/cvss/830.html.
I have accomplished the two approaches that we discussed. As I have presented, for approach 1, we choose ten men and ten women for the training phase, to get a strong encoder. Then we fix the encoder and finetune the decoder, expecting to train the style of the target speaker. For approach 2, we train 4 exemplar autoencoders with speech from 4 speakers, then we fix all the decoders and train a public encoder for content extraction. Finally, we train a decoder for the target speaker to capture speaker style.
It feels like that Approach 2 does present a better performance in cross-gender task.

Timeline

Still not good enough

2021.9.13～2021.10.20　　consider introducing *loop cor*



2021.9.20　　introduce multi-step training，use griffin-lim as vocoder for training phase；after this step，Fix this model and train the wavenet vocoder

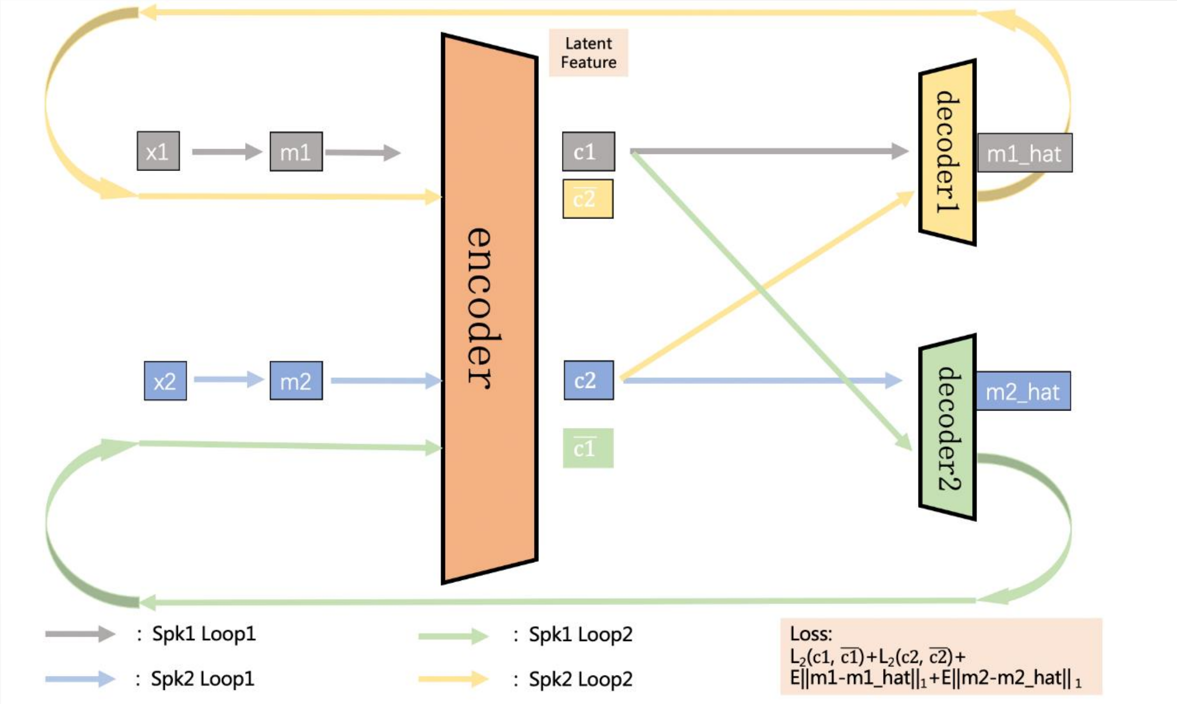
How to prove a better encoder → Check the content code!

2021.9.29～2021.10.10　　use Tsne to observe the clustering ability of content code，and decide a best encoder

2021.11.10　　report on **Cycle-Loss based Exemplar Autoencoder**

Cycle loss based Exemplar Encoder

- **1st round encoding**: Firstly convert x1 and x2 into spectrum m1 and m2; encode into latent space. Save latent features as c1 and c2.
- **Speech reconstruction**: Construct two decoders specific to speaker s1 and s2. Forward c1 and c2 to the decoder and produce the reconstructed spectrum m1_hat and m2_hat.
- **2nd round encoding**: Forward c1 and c2 separate to decoder2 and decoder1; then encode through common encoder again for latent features $\overline{c1}$ and $\overline{c2}$



Loss:

$$L_{cycle} = L_2(c1, \overline{c1}) + L_2(c2, \overline{c2})$$

$$L_{spec} = E||m1 - m1_{hat}||_1 + E||m2 - m2_{hat}||_1$$

$$L = \alpha * L_{cycle} + L_{spec}$$

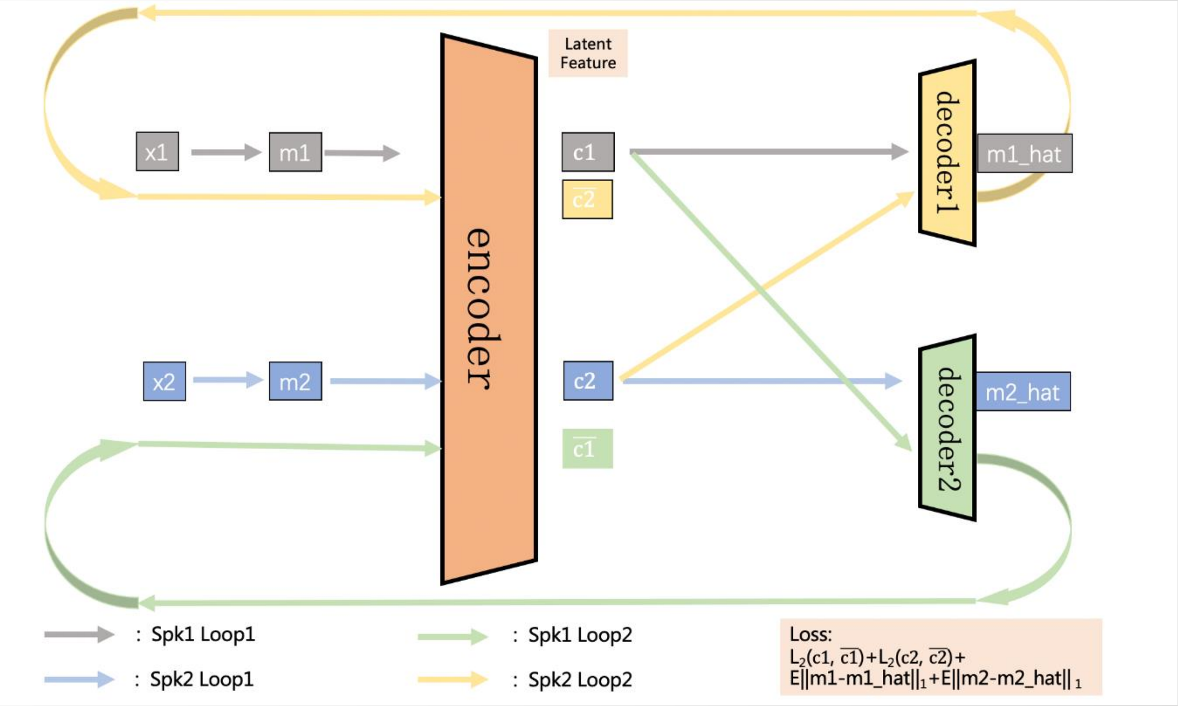
Multi-Step Training

- **1st step**: Introduce cycle loss for a stronger encoder.

Loss:

$$L_{cycle} = L_2(c1, \overline{c1}) + L_2(c2, \overline{c2})$$
$$L_{spec} = E||m1 - m1_{hat}||_1 + E||m2 - m2_{hat}||_1$$
$$L = \alpha * L_{cycle} + L_{spec}$$

- **2nd step**: Fix the encoder and finetune the decoder for an autoencoder for a specific speaker.



Check latent code to verify a best encoder

- We extract the content code from the output of the encoder and use this code for a further test.
- First, we choose six phones from the same speaker of the training period, each of which consists of 6 samples.
- Then set these phones as input into the autoencoder, and we can get the latent codes of these phones.
- Use tSNE to observe the clustering capibility of the phones. The dimension of the output of TSNE is 2.

Timeline

How to prove that cycle loss is useful?

2021.10.20～2021.11.3 Multiple Tasks

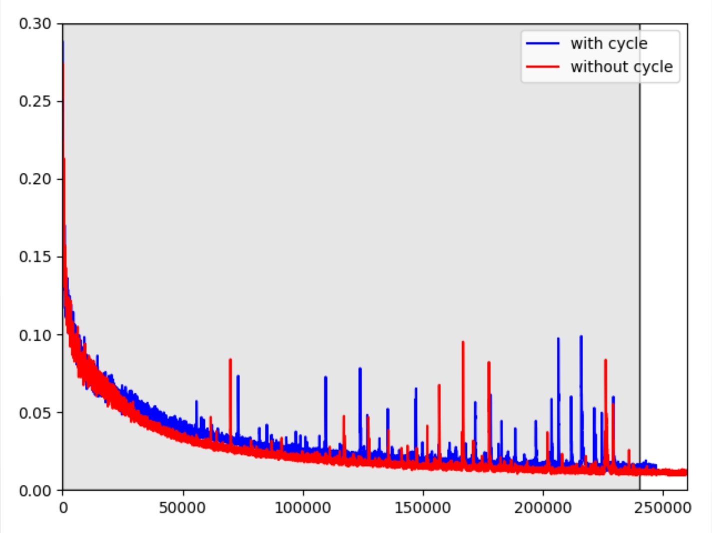
- Test of comparison between cycle-loss model and multi-decoder model without cycle loss
- Test of comparison between different IB dimensions

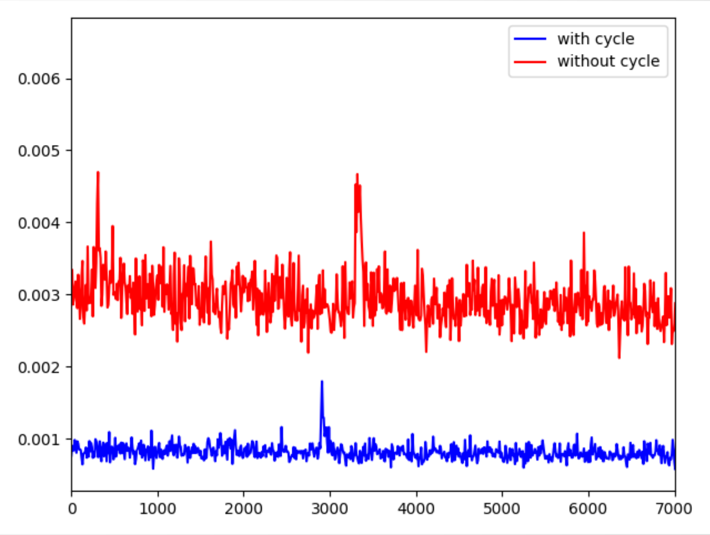
2021.11.9～2021.11.11 Qualitive Tests and Website update

2021.11.12～2021.11.15 Loss curve *How quantitative?*

2021.11.18 GOP & SCA tools ready

2021.11.23 GOP test





Timeline

2021.11.25 Review other recent improvements

- FRAGMENTVC: ANY-TO-ANY VOICE CONVERSION BY END-TO-END EXTRACTING AND FUSING FINE-GRAINED VOICE FRAGMENTS WITH ATTENTION

- ANY-TO-ONE SEQUENCE-TO-SEQUENCE VOICE CONVERSION USING SELF-SUPERVISED DISCRETE SPEECH REPRESENTATIONS

What about Wav2vec + Decoder?

They use wav2vec to sequence to train any-to-one.

Timeline

2021.12.5 Finish a first draft of the paper; a new thinking *on Never-before-seen Speaker Conversion,* with simple fine-tune on decoder and modify to a new style for conversion while fixing the encoder

2021.12.8～2021.12.20 GOP test on *Never-before-seen Speaker Conversion* task

2021.12.24～2021.12.28 Submit a patent

2022.1.9～2022.1.24

- Paper sharing on VQW2V
- VQW2V and cycle+VQW2V model configuration and training
- Paper modifying

Timeline

2022.1.28 Review paper on CycleVAE

- MANY-TO-MANY VOICE CONVERSION USING CYCLE- CONSISTENT VARIATIONAL AUTOENCODER WITH MULTIPLE DECODERS

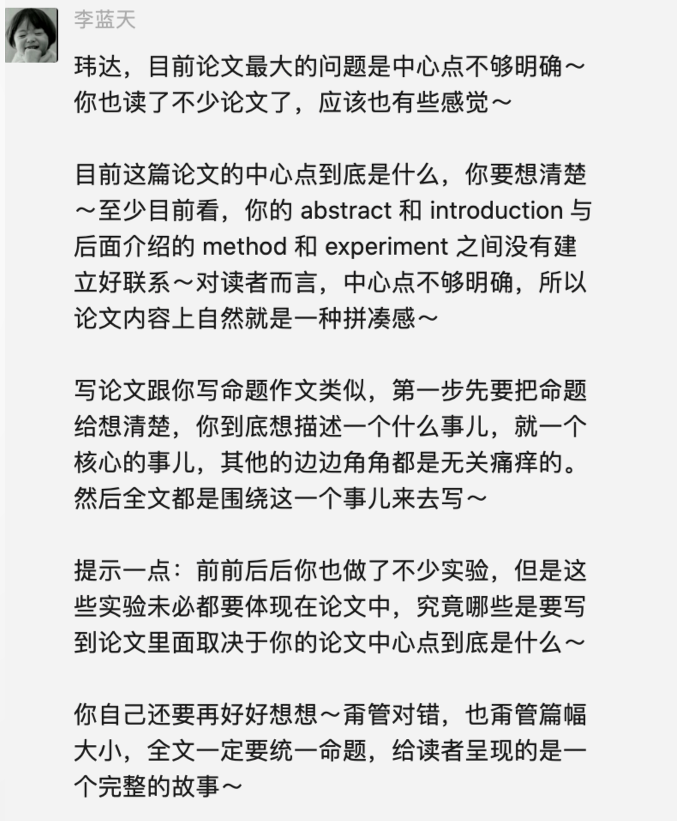
Cycle on code or spk?

2022.1.24～2022.2.16

- Paper and patent modification
- Add CycleVAE comparison results
- Interspeech Paper Reading

2022.2.24～2022.3.14

- Finish all experiments for the paper



李蓝天

玮达，目前论文最大的问题是中心点不够明确～你也读了不少论文了，应该也有些感觉～

目前这篇论文的中心点到底是什么，你要想清楚～至少目前看，你的 abstract 和 introduction 与后面介绍的 method 和 experiment 之间没有建立好联系～对读者而言，中心点不够明确，所以论文内容上自然就是一种拼凑感～

写论文跟你写命题作文类似，第一步先要把命题给想清楚，你到底想描述一个什么事儿，就一个核心的事儿，其他的边边角角都是无关痛痒的。然后全文都是围绕这一个事儿来去写～

提示一点：前前后后你也做了不少实验，但是这些实验未必都要体现在论文中，究竟哪些是要写到论文里面取决于你的论文中心点到底是什么～

你自己还要再好好想想～甭管对错，也甭管篇幅大小，全文一定要统一命题，给读者呈现的是一个完整的故事～

Timeline

2022.3.21 Submit abstract for Interspeech Paper

Still lack theoretical analysis…

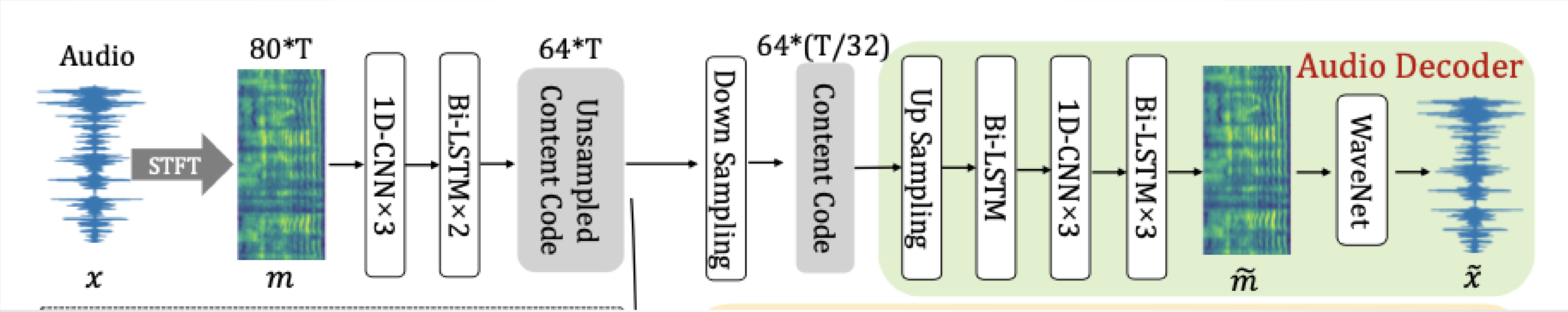
2022.3.23～2022.3.27 Add supplementation tests

- Comparison with AutoVC
- UMAP on word/phone level clustering, for theoretical analysis

2022.3.28 Submit the final paper

2022.4.3～2022.4.8 Submit code and modify project website

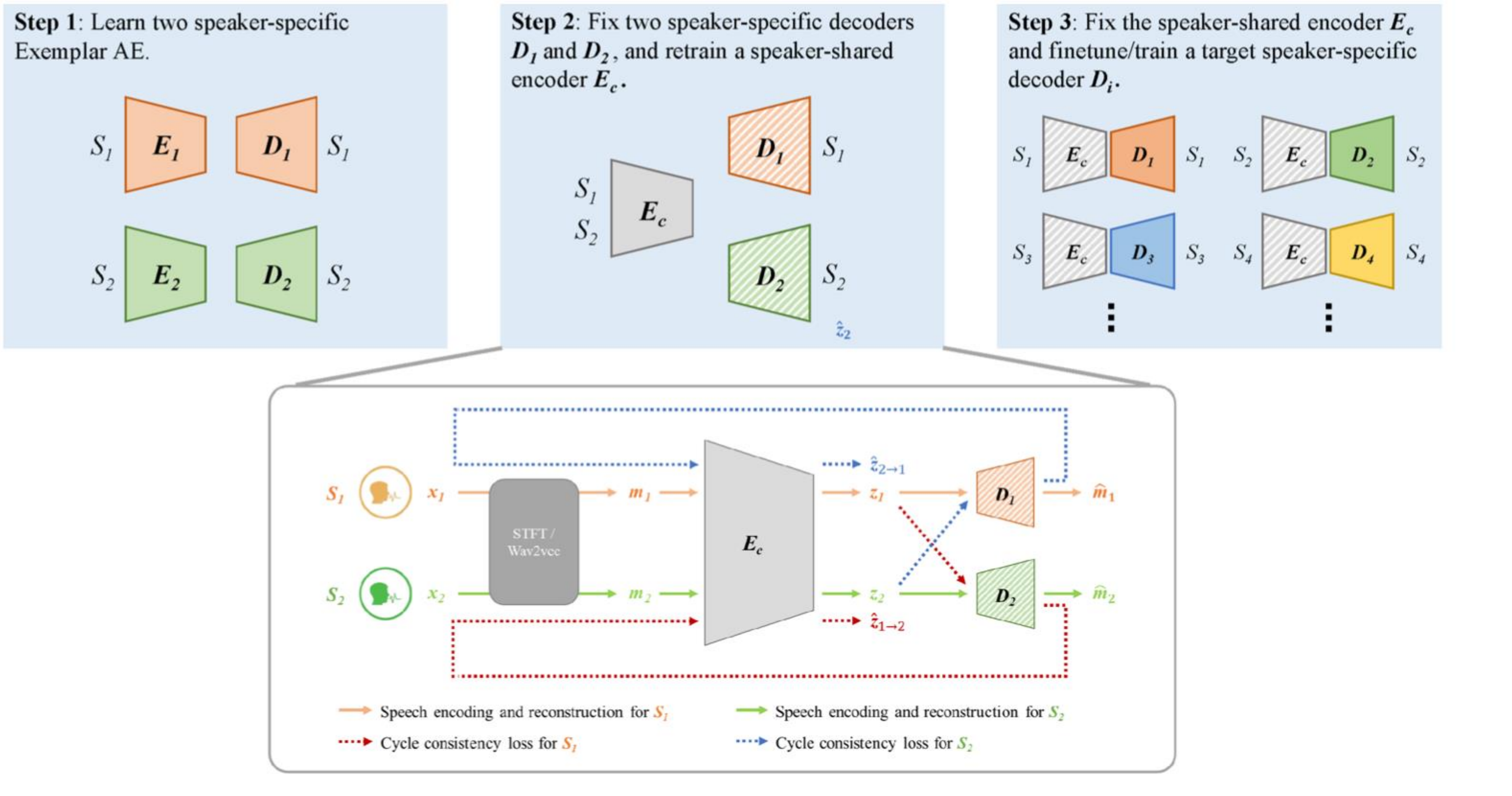
Exemplar Autoencoder



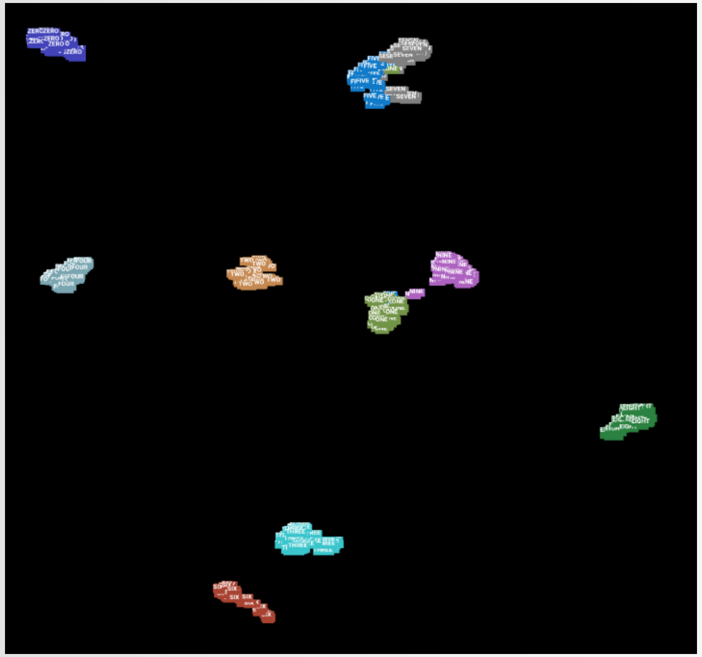
Encoder **Decoder** **Vocoder**

Kangle Deng, Aayush Bansal, Deva Ramanan, “UNSUPERVISED AUDIOVISUAL SYNTHESIS VIA EXEMPLAR AUTOENCODERS“ in ICLR 2021

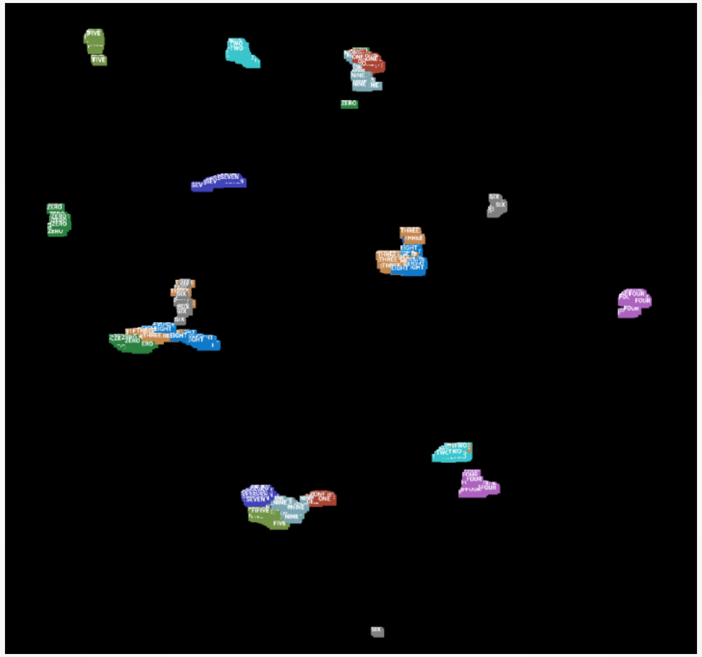
Enhanced exemplar autoencoder



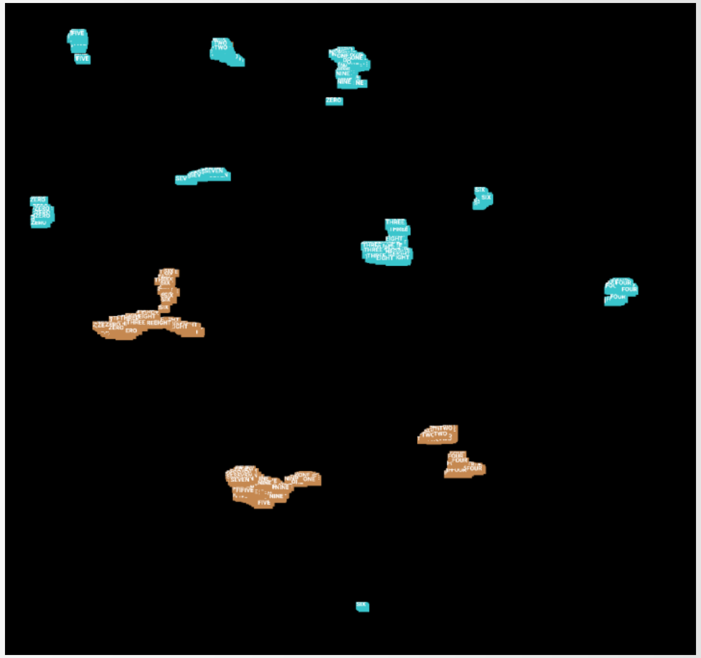
Theoretical Analysis with UMAP



单人UMAP



一男一女UMAP(按内容)

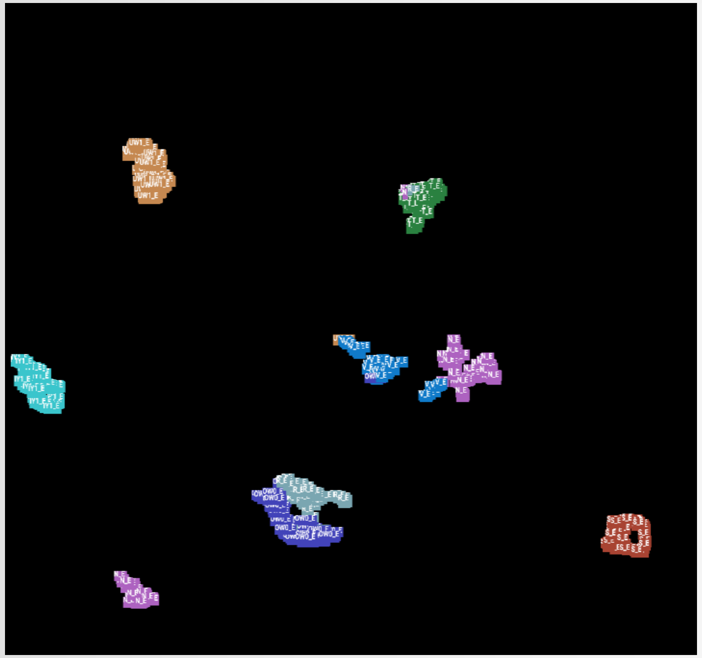


一男一女UMAP(按性别)

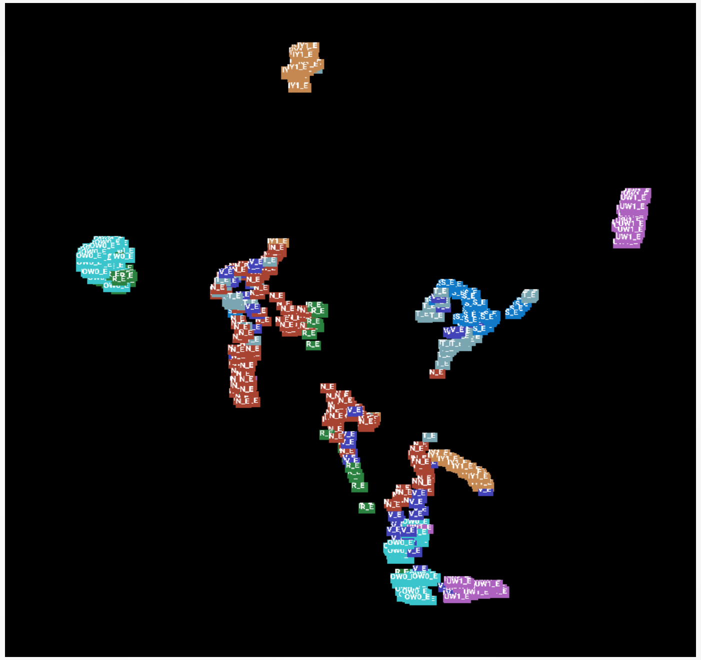
Word Level

Speaker variation is more significant than content variation.

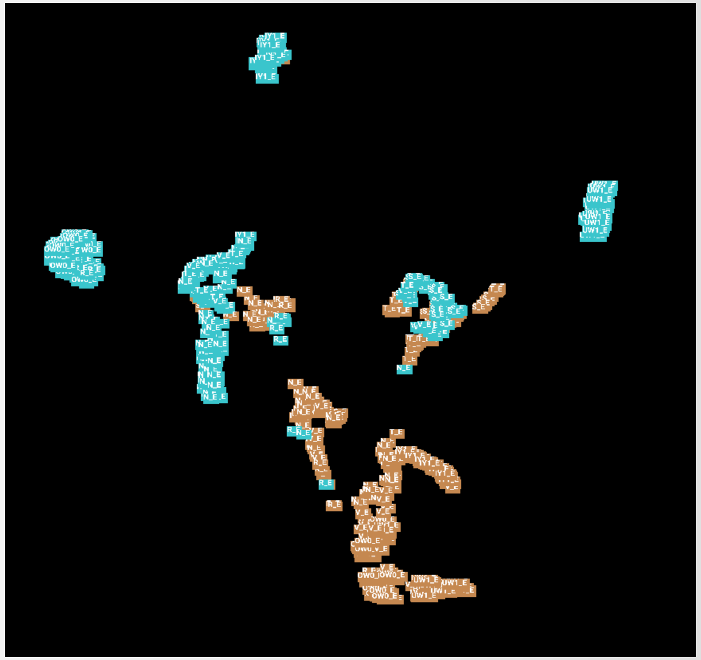
Theoretical Analysis with UMAP



单人UMAP



一男一女UMAP(按内容)



一男一女UMAP(按性别)

Phone Level

Speaker variation is more significant than content variation.

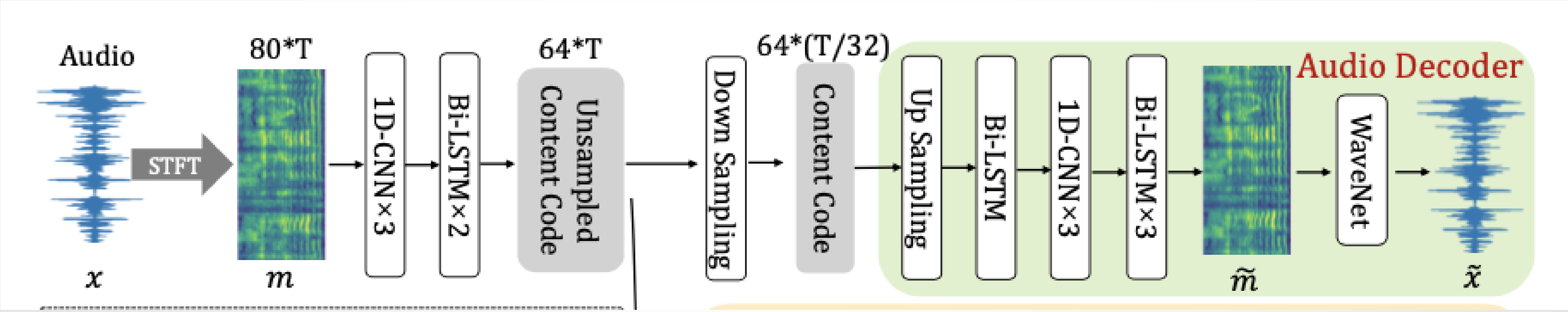
Dataset, model and metrics

- speech data from the AISHELL-3 dataset
- All the speech signals are formatted with 16kHz sampling rate and 16-bits precision.
- No overlap in speakers exists between the training and test sets.

Table 1: *Data profile*

Set	# of Spks	Utters per Spk	Duration per Spk
Train	4 (2 Female, 2 Male)	∼400	∼25 mins
Test	6 (3 Female, 3 Male)	∼250	∼15 mins

Model



Encoder **Decoder** **Vocoder**

Metrics

- GOP, CER, SCA and MOSNet
    - GOP and MOSNet primarily evaluate the quality of the generation
    - CER mostly focuses on intelligibility
    - SCA is more related to resemblance to the target speaker
- The Kaldi toolkit is used to compute CER and GOP.
- A pre-trained model is used to predict the MOSNet score.
- For SCA test, we train a speaker classification model based on the x-vector structure with 400 background speakers from AISHELL-1 dataset plus the target speakers from the training set.

Main Results

- Same-gender case

- Cross-gender case

Table 2: *Comparison between eAEs with/without cycle consistency loss. SG and CG denote the same-gender and cross-gender tests respectively.*

		GOP (↑)	**CER(%)** (↓)	**MOSNet** (↑)	**SCA(%)** (↑)
eAE	SG	1.489	19.29	2.712	81.85
	CG	1.368	21.19	2.668	80.00
eAE + Cycle	SG	1.605	14.27	2.786	85.00
	CG	1.589	14.19	2.778	85.45

Generalization to new target speakers

In this test, we firstly train an eAE with cycle consistency loss as in the previous experiment, and then fix the encoder and train decoders for 6 new speakers selected from AISHELL-3.

The same test data in the test set are used to perform test on these new target speakers. For comparison, we also train 6 individual vanilla eAEs for the same 6 speakers.

Table 3: *Performance on new target speakers.*

	GOP (↑)	**CER(%)** (↓)	**MOSNet** (↑)
eAE	1.439	20.86	2.718
eAE + Cycle	**1.539**	**15.23**	**2.760**

Ablation Study

- More Training Speakers
  - 1 vs 2 vs 4
- Code cycle and data cycle
- Encoder sharing or cycle loss
- Work with powerful front end

Table 4: *Results of ablation study.*

No.	Model	# Spks	GOP	CER(%)	MOSNet	SCA(%)
1	eAE	1	1.368	21.19	2.768	80.00
2	eAE + Cycle	2	1.589	14.19	2.778	85.45
3	eAE + Cycle	4	1.593	14.03	2.737	85.10
4	eAE + En-Share	2	1.378	21.28	2.689	80.40
5	eAE + Data Cycle	2	1.513	18.56	2.724	82.80
6	eAE/W2V	2	1.612	11.88	2.795	89.25
7	eAE/W2V + Cycle	2	1.713	10.73	2.823	89.60

Conclusion

- In this paper, we proposed an enhanced exemplar autoencoder for any-to-one voice conversion.

- The core design is a cycle consistency loss, which enforces the content code of the recon- structed speech close to the original speech, no matter by whose decoder decodes the speech.

- We demonstrated theoretically and empirically that the proposed technique can significantly purify the content code, and produce better performance in complex VC tasks, such as cross-gender conversion.

Others

- Some feelings for doing researches
  - Work on your own first before asking others' help
  - Update to your mentor in time when meeting problems
  - Keep your own rhythm and self-push
  - Get used to facing problems
  - Always make your work better and more convincing
  - Schedule and plan first before doing tasks

Others

- Some useful tools
  - Make your plans: 幕布、石墨文档、notion …
  - 画图: PPT、Embedding Projector(Google)
  - Study via: bilibili、CSDN、Google Scholar、知网硕士论文…
  - Paper reviewer: Endnotes …
  - Update your status: Weekly meeting、CVSS

Thank you!